SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
GENOME
POLYPLOIDY
PLANT GENETICS
GENETICS

Received
15 May 2013

Accepted
15 November 2013

Published
2 December 2013

SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids

Antonio A. F. Garcia[1]*, Marcelo Mollinari[1]*, Thiago G. Marconi[1,2], Oliver R. Serang[3], Renato R. Silva[1], Maria L. C. Vieira[1], Renato Vicentini[2], Estela A. Costa[2], Melina C. Mancini[2], Melissa O. S. Garcia[2], Maria M. Pastina[1], Rodrigo Gazaffi[1], Eliana R. F. Martins[4], Nair Dahmer[4], Danilo A. Sforça[2], Claudio B. C. Silva[2], Peter Bundock[5], Robert J. Henry[6], Glaucia M. Souza[7], Marie-Anne van Sluys[8], Marcos G. A. Landell[9], Monalisa S. Carneiro[10], Michel A. G. Vincentz[2,4], Luciana R. Pinto[9], Roland Vencovsky[1] & Anete P. Souza[2,4]

[1]Departamento de Genética, Escola Superior de Agricultura ''Luiz de Queiroz'' Universidade de São Paulo, Brazil, [2]Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, SP, Brazil, [3]Department of Neurobiology, Harvard Medical School and Proteomics Center, Children's Hospital Boston, USA, [4]Departamento de Biologia Vegetal, Instituto de Biologia, Universidade Estadual de Campinas, Campinas, SP, Brazil, [5]Southern Cross University, Lismore, Australia, [6]Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Brisbane, Australia, [7]Instituto de Química, Universidade de São Paulo, São Paulo, SP, Brazil, [8]Departamento de Botânica, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP, Brazil, [9]Centro de Cana, Instituto Agronômico de Campinas, Ribeirão Preto, SP, Brazil, [10]Departamento de Biotecnologia de Plantas, Centro de Ciências Agrárias, Universidade Federal de São Carlos, São Carlos, SP, Brazil.

Correspondence and requests for materials should be addressed to A.P.S. (anete@unicamp.br)

* These authors contributed equally to this work.

Many plant species of great economic value (e.g., potato, wheat, cotton, and sugarcane) are polyploids. Despite the essential roles of autopolyploid plants in human activities, our genetic understanding of these species is still poor. Recent progress in instrumentation and biochemical manipulation has led to the accumulation of an incredible amount of genomic data. In this study, we demonstrate for the first time a successful genetic analysis in a highly polyploid genome (sugarcane) by the quantitative analysis of single-nucleotide polymorphism (SNP) allelic dosage and the application of a new data analysis framework. This study provides a better understanding of autopolyploid genomic structure and is a sound basis for genetic studies. The proposed methods can be employed to analyse the genome of any autopolyploid and will permit the future development of high-quality genetic maps to assist in the assembly of reference genome sequences for polyploid species.

Common marker systems, such as Amplified Fragment Length Polymorphism (AFLP) and Simple Sequence Repeat (SSR), have been successfully used in the last few decades for several types of genetic studies, including diversity analysis, genetic mapping, quantitative trait locus (QTL) mapping, synteny (co-linearity) definition, co-ancestry estimation, and more. However, most of these applications have been developed in diploid plant species in which the theoretical foundation for analysis and interpretation of the results has already been established. These tools are less developed for autopolyploids, i.e., organisms that have more than two sets of chromosomes of the same type and origin[1]. Despite the fact that great progress has been made using marker systems in autotetraploids (e.g., potato), other, more complex polyploid species, such as sugarcane, strawberry, and some forage crops, have not yet fully benefited from molecular marker information.

This is because several unrealistic and simplified assumptions need to be made. AFLP and SSR (and even RFLP) do not allow a straightforward estimation of the number of copies of each allele (dosage) at a given polymorphic locus in complex polyploids (species with more than four chromosomes per homology group). For example, in sugarcane, there are approximately 22 linkage maps[2], and only a few of these maps include loci with high allelic doses. The scenario is similar for QTL studies[3]. Some models have attempted to consider the effects of QTL dosage[4,5], but these models still rely on marker data that are not fully informative. In *Saccharomyces cerevisiae*,

microarray studies have demonstrated that gene expression and gene regulation may depend upon the ploidy level[6], emphasising that allele dosage should be included in marker-assisted selection or genome-wide association studies. Similar conclusions were reached by[7].

The development of modern genotyping technologies, such as the Sequenom iPLEX MassARRAY®[8], Illumina GoldenGate™[9], and protocols such as Genotyping by Sequencing (GBS[10]) or RAD seq[11], allows the evaluation of single-nucleotide polymorphisms (SNP) throughout the genome. One interesting feature of these novel approaches is the possibility of evaluating the relative abundance of each allele, i.e., the allelic dosage. This significantly increases the information embodied in each locus and provides several advantages for genetic analysis, such as mapping mutants via quantitative bulked segregant analysis[12] and the possibility of estimating ploidy level for polyploids[13,14]. For complex polyploids, such as sugarcane, this is essential because each marker locus needs to be positioned in a homology group. What is remarkable is that the sugarcane homology groups have different numbers of chromosomes[15]. This makes the estimation of ploidy level for each SNP an essential step for further analysis. Furthermore, less studied polyploid species with unknown ploidy levels could directly benefit from this modern marker approach.

To illustrate one of the advantages of using SNPs for these purposes, let us assume a hypothetical population of an autohexaploid species having the following genotypes for a given locus: *aaaaaa*, *Aaaaaa*, *AAaaaa*, *AAAaaa*, and so on up to *AAAAAA*. Using the *A* allele as reference, these individuals are said to have between zero (nulliplex) and six copies (hexaplex) of the allele. The number of copies of the reference allele is the allele dosage. If the individuals are evaluated with a marker system, such as AFLPs or SSRs, they are scored as 0 (gel band absent) for *aaaaaa* or 1 (gel band present) for all the other individuals due to one intrinsic limitation of the method that is associated with overlooking ploidy level. Thus, a result of "1" in a binary marker system indicates the presence of at least one copy of allele A. However, if SNPs are evaluated, the scores will be *0A* : *6a*, *1A* : *5a*, *2A* : *4a*, and so on up to *6A* : *0a* (this allelic dosage notation will be used throughout this manuscript). A marker system that allows for the direct observation of all genotypes is therefore much more informative and should be preferred. Nevertheless, this raises new challenges because new statistical methods must be developed to allow for the comprehensive analysis and interpretation of data in this new scenario.

In this work, we have evaluated the use of SNPs and novel statistical methods for SNP calling and ploidy level estimation in sugarcane using mass spectrometry-based procedures and the SuperMASSA software[13,14]. We demonstrate that it is possible to estimate the ploidy level and the dosage of SNPs, providing useful insights into the sugarcane genome interpretation. Sugarcane is an excellent test case because it is a complex polyploid with an unknown ploidy level and frequent aneuploidy[15]. This work will make studies on linkage and QTL mapping, association mapping, and genomic selection possible by bringing the advantages of molecular markers to complex polyploids that, with the exception of a few well-studied autotetraploids (such as potato), have poorly understood genomes. We explored two different scenarios. First, 271 SNPs generated using the Sequenom iPLEX MassARRAY technology[8] were used to analyse a population of 180 individuals from a biparental cross between the varieties IACSP95-3018 and IACSP93-3046. Second, 1034 SNPs were analysed in a panel of 142 relevant sugarcane genotypes. The panel consisted of important commercial varieties in addition to ancestral and parental genotypes that have been frequently used in a wide spectrum of breeding programs.

Results

Figure 1, panels A.1, B.1, and C.1 show examples of scatter plots of genotypes in the segregating population for a selected SNP (*SugSNP382*). It is clear that there are three clusters of points, each corresponding to one genotype. The data are shown together with dotted lines indicating the expected angles where the individuals would be placed if the ploidy level were 8, 10, and 12. The results suggest that the ploidy level was 10 because the clouds of points deviated slightly from the lines to other ploidy levels. When observing the data from the parents (Figure 1, A.2, B.2, C.2) and considering the closest distance to the expected genotype, the deduced configurations should be $8T:0G \times 6T:2G$, $10T:0G \times 8T:2G$, and $12T:0G \times 9T:3G$. We must note that to be consistent with the number of observed clusters (three), the expected genotype distributions in the population were set to assume that the locus had a double dosage in one parent and was nulliplex in the other. The deduced value for ploidy 12 was not consistent with the putative number of observed clusters (three) in the progeny because a triple-dosage locus would allow for four clusters in the progeny. The expected population ratios (Figure 1, A.3, B.3, C.3) were slightly different for each ploidy level, with 3 : 8 : 3, 2 : 5 : 2, and 5 : 12 : 5 values for octa-, deca-, and dodecaploidy, respectively. It must be emphasised that it would be extremely difficult to distinguish these levels only by inspection or even by a simple statistical test with reasonable sample sizes.

The results described above help to explain the complex scenarios involved in determining ploidy and dosage. These issues have recently been analysed using the statistical procedures included in the SuperMASSA software[14]. The model simultaneously considers all available information and the genetic constraints that the derived results must fulfil, i.e., the possible genotypes to be observed given the ploidy level and the parental genotypes, the ratio between allele intensities, and the expected complete polysomic segregations. This allowed the exclusion of a triple dosage for ploidy 12. Because the expected segregations are similar, the classification relies on the ratio of the alleles (indicated by dotted lines on Figure 1), and this is one of the reasons why the choice of a technology with less bias for ratios is essential. These issues have been thoroughly discussed in[14]. Those authors analysed how to address situations where some bias is present. In our previous experience with Sequenom and Illumina data[13], we observed that the former experimental approach is much less likely to produce an allele ratio bias.

We present a deeper analysis of SNPs using SuperMASSA[14] in Figure 2, where the statistical results for three selected SNPs are depicted. For *SugSNP382* (described previously in Figure 1), the results indicate that the *posterior* probability of ploidy 10 is close to 1; all individuals were allocated to clusters with individual *posterior* probabilities no smaller than 0.6 (almost all these probability values were close to 0.9). There was also a good agreement between the observed and expected distribution of the genotypes in the biparental population. In addition, we can deduce that the parental genotypes must have been $8T:2G \times 10T:0G$. The preliminary visual inspection of the scatter plot described in Figure 1 is consistent with our statistical results.

For *SugSNP151* and *SugSNP715*, the other two SNPs shown in Figure 2, the analysis is more complicated. Although it was possible to find models with high *posterior* probability for ploidy levels 18 and 16, the individual *posterior* probabilities in both cases were all smaller than 0.6. This means that if a small naive *posterior* threshold of 0.65 were used, none of the individuals would be classified as having a specific genotype. This clearly shows that, as reported previously[14], the *posterior* probability cannot be used as a single criterion to interpret the results. There were also differences between the observed and expected distributions. Although this result may not be considered reliable enough to interpret the available laboratory data, the most likely configuration for these SNPs is ploidy 18 and 16, with parental genotypes of $15G:3A \times 12G:6A$ and $10T:6C \times 7T:9C$, respectively.

The estimates of ploidy level for the 249 SNPs evaluated in the biparental population fell between 2 and 20 (Figure 3a). An examination of three loci classified as having a ploidy of 2 (*SugSNP_0004*, *SugSNP_0033* and *SugSNP_0036*) showed that these results are

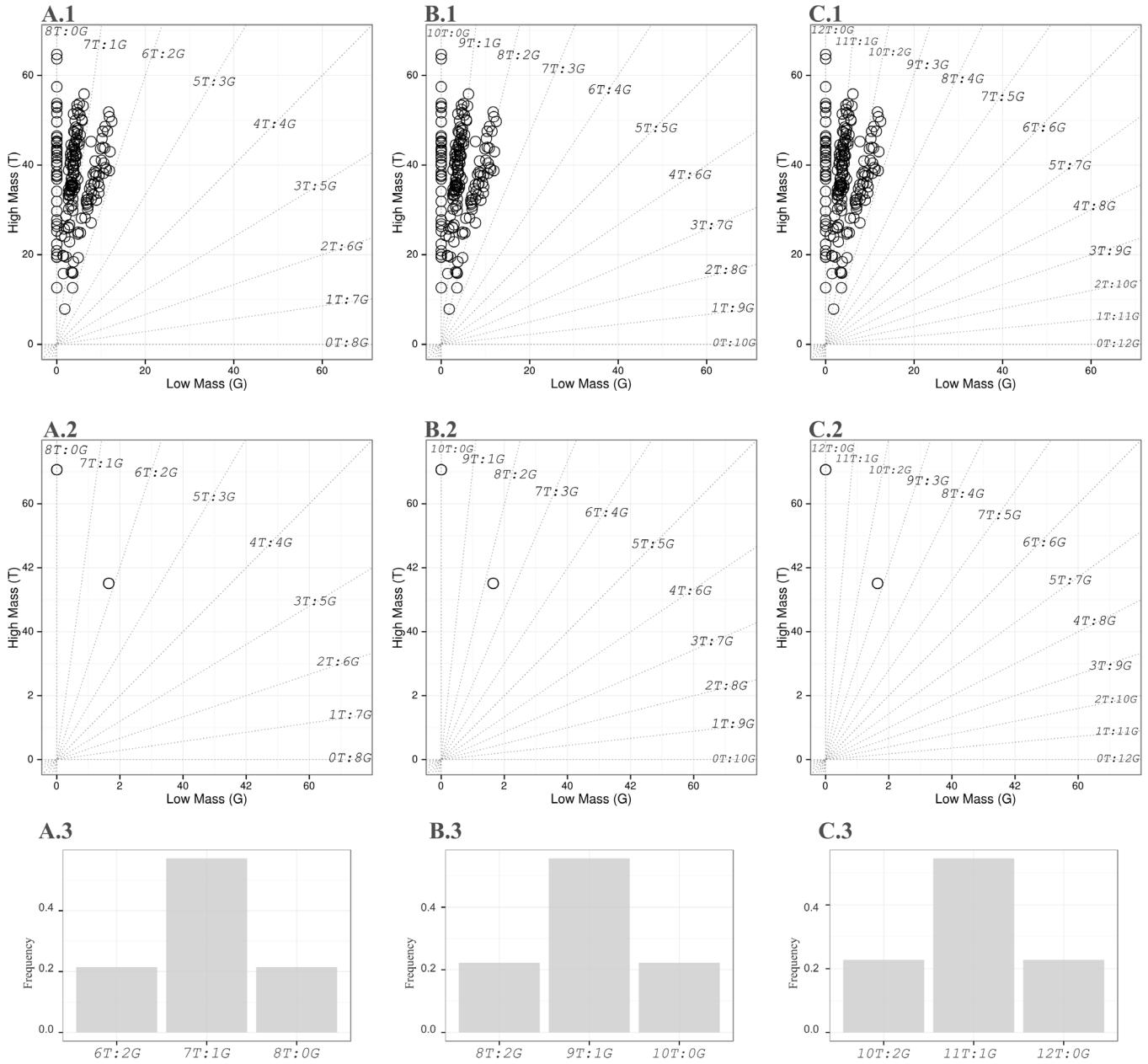

Figure 1 | A panel of nine graphs, with three columns (A, B, and C) representing ploidy levels of 8, 10, and 12, respectively. Row 1: Raw data of two allele intensities for *SugSNP382* in the biparental segregating population. Dotted lines show the possible genotypes and the allele ratios that could be observed for each corresponding ploidy level. Row 2: allele intensities for the parents of the population (the average of 12 replicates), also considering the respective ploidy level. Row 3: expected segregations for the respective ploidy level and assuming parents with genotypes $8T{:}\,0G \times 6T{:}\,2G$, $10T{:}\,0G \times 8T{:}\,2G$, $12T{:}\,0G \times 10T{:}\,2G$; these genotypes were chosen based on a visual inspection of rows 1 and 2.

clearly associated with data of poor quality. The ratios between the masses of these alleles did not follow any expected pattern and were quite different from what was observed for all other SNPs. Therefore, these SNPs were not included in the final presentation of the results (Figure 3); for the same reason, five loci with ploidy 4 were also discarded (*SugSNP_0011*, *SugSNP_0017*, *SugSNP_0018*, *SugSNP_0048* and *SugSNP_0083*); note that another two SNPs with ploidy 4 (*SugSNP_0008* and *SugSNP_0061*) are presented in Figure 3; both had a single allelic dosage in one of the parents.

The procedure to develop the SNPs must not, in principle, exclude or favour any homology group. In our analysis, only 2 out of 249 loci were classified as having a ploidy of 4 and a single dosage, but there are no reports of such ploidy levels in the sugarcane literature. We must conclude that it is unlikely that sugarcane has homology groups with four chromosomes (autotetraploid). One possible explanation is that the observed results were caused by some bias in the angles of the scatter plots. If the PCR amplification has a different efficiency for each chromosome, the ratio between the allele intensities may be slightly different from the real ratio and therefore the angles of lines in the scatter plots could be biased by these differences (please see the additional simulations examining this bias in the Supplementary Material). As explained for Figure 1, for small dosages, the differences in the expected segregations are virtually indistinguishable and rely heavily on the scattered plot angle estimation; therefore, if this bias was present, some loci may have been misclassified as autotetraploids. We applied the same reasoning when analysing the association mapping panel; consequently, loci with an estimated ploidy of 2 or 4 were not included in the final results (Figure 3b), and of the 987

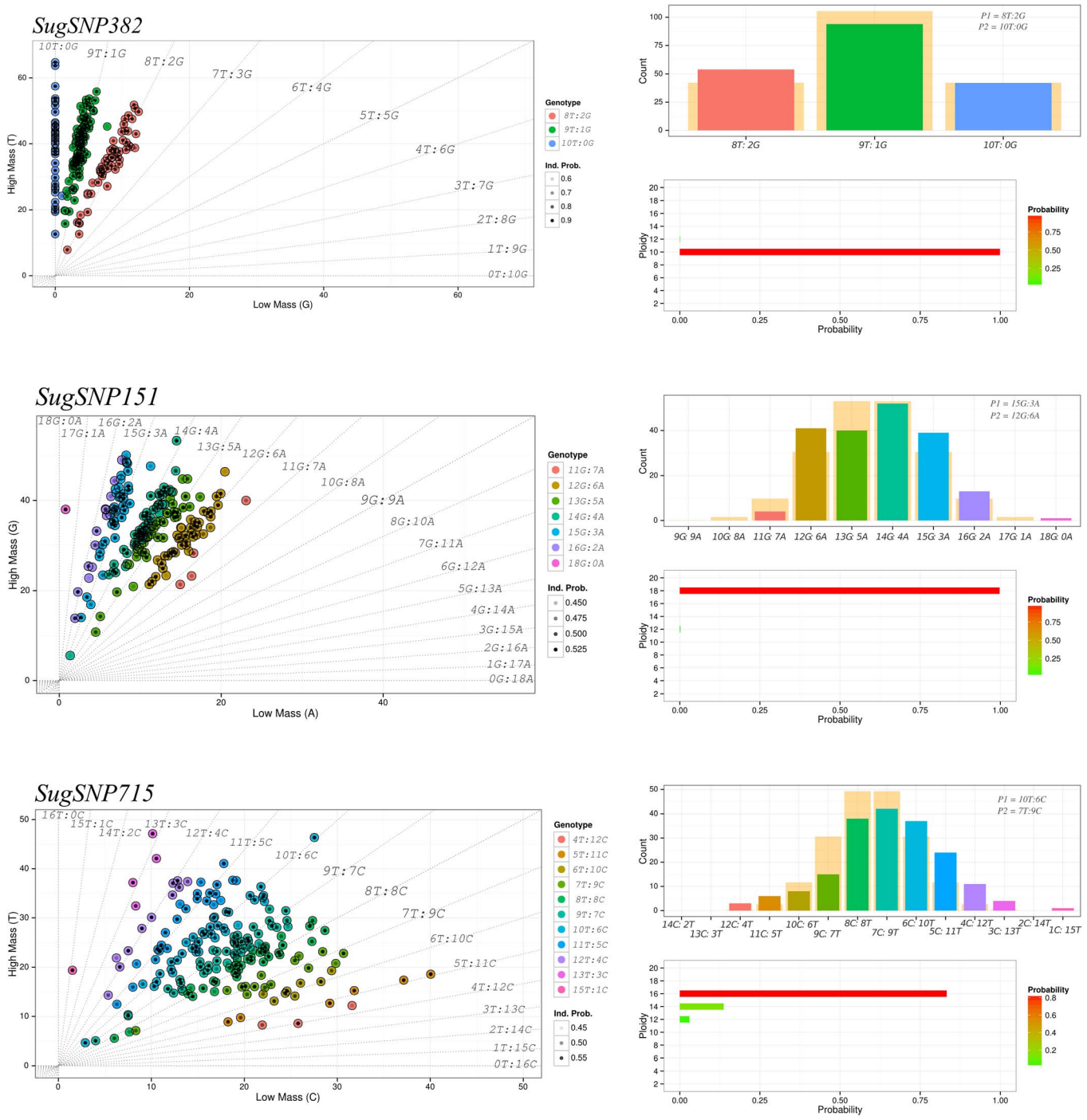

Figure 2 | Results of statistical analysis for three selected SNPs in a biparental sugarcane population. Each panel of three graphs correlates to one SNP. The scatter plots show the classification of each individual in a cluster (genotype), indicated with a different colour; the centre of each circle has a small grey dot, whose colour intensity indicates its posterior probability of being allocated in the cluster. Expected (in yellow) and observed distributions for the estimated ploidy level and dosage on the parents are indicated on the histograms; the same colours used on the scatter plots were considered for the observed distribution. The *posterior* probabilities for each ploidy evaluated in the range 2 to 20 (only even numbers) are also indicated.

SNPs that were initially available (after quality control), 855 were taken into account. For all other ploidy levels, the number of loci within each ploidy class suggests that our results are reliable. The ploidy levels fell between 6 and 20, showing that the number of chromosomes within the homologous groups is not constant in sugarcane, which is in agreement with previous results[15].

The distribution of loci within each ploidy level and category (A, B, and C) was similar for both the biparental population (Figure 3a) and the panel of sugarcane genotypes (Figure 3b), with the exception of those loci with ploidy 20, which were more frequent in the panel. All of the category A ploidy levels seemed to be present in about the same proportions (except ploidy 4, which was likely to be a misclassification) in both scenarios (Figures 3a and 3b). For category B, there was a trend of having more loci with higher ploidy levels; this was even clearer for loci of category C, particularly for the biparental population (Figure 3a), where none of the loci had a ploidy level smaller than 12.

It is important to mention that the analysis of the 142 sugarcane genotypes within the panel (Figure 3a) was much more complicated

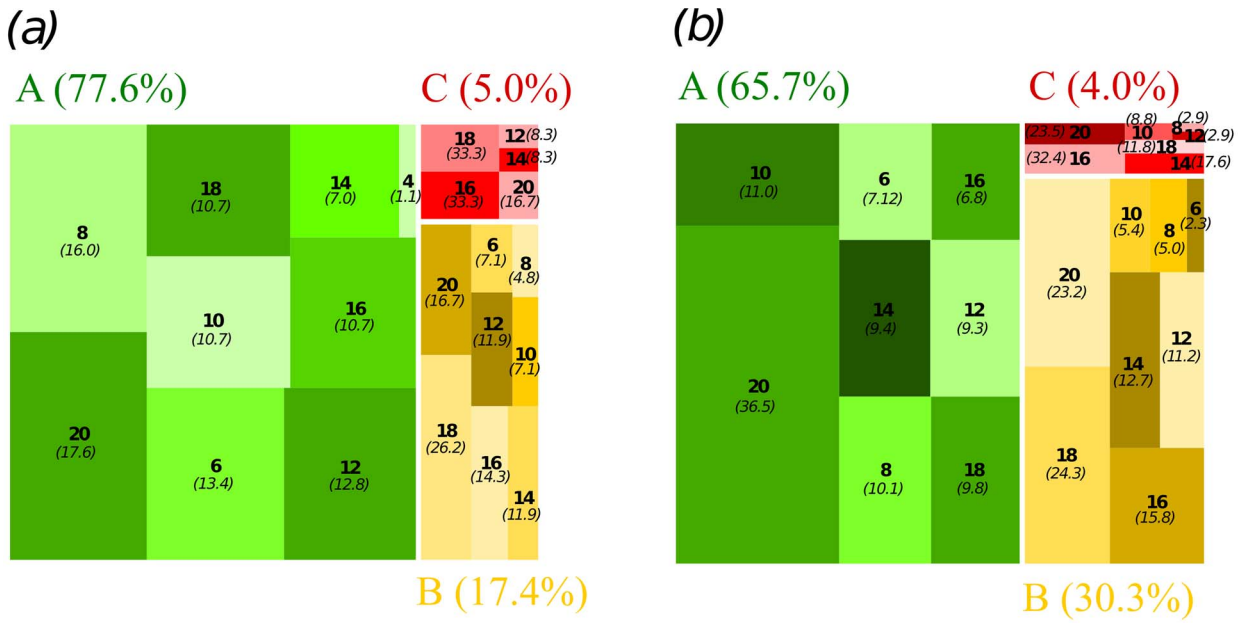

Figure 3 | Representation of the estimates of ploidy level (in bold font) for the configurations with highest posterior probabilities for the biparental population (a) and association mapping panel (b). The areas of the rectangles are proportional to the number of SNPs that have the same ploidy level, indicated within each rectangle in parenthesis. According to the posterior probabilities calculated for each even-numbered ploidy level in the range 2 to 20, each SNP was classified into one category, using the following *ad hoc* criteria: Category A (green), when the highest posterior probability is greater than or equal to 0.80; Category B (yellow), when no single value of the posterior probability is higher than 0.80 but the sum of the two highest ones is greater than or equal to 0.80; and Category C (red): all other cases. In parentheses: the number of SNPs within the given ploidy level and category. The total SNP number for (a) was 241, and the total SNP number for (b) was 855.

because there were no parental genotypes available to guide the analysis, as there were in the biparental population. For this group, we assumed Hardy-Weinberg equilibrium and that all individuals had the same ploidy level for a given locus. Given the complexity of the sugarcane genome, this assumption may seem rather strong, but the final genetic results are consistent with a ploidy level distribution similar to that of the biparental population. Again, we observed that the number of chromosomes within homologous groups was not constant.

The number of single-dose loci for SNPs in categories A and B (Figure 4) indicates that these are more frequent for ploidy levels up to 12. At the ploidy level of 20, only 5% of the SNPs were single dose. It is remarkable that so few SNPs overall had single-dose alleles. Interestingly, if these SNPs were used to build a linkage map using the conventional approach (single-dose markers), only loci classified as single-dose loci in IACSP93-3046 and as nulliplex in IACSP95-3018 (or vice-versa) or those classified as single-dose loci in both parents would be considered. Only 76 (30.5%) SNPs would meet these criteria if all ploidy levels were considered altogether.

The results presented in Figures 3 and 4 are interesting and informative, but because they are based only on the *posterior* probability of a ploidy level, they need to be interpreted together with individual probabilities[14]. Figure 5 shows that the analysis of ploidy levels 6, 8, and 10 was more reliable, as most of the loci had medians for the individual *posterior* probabilities in the range 0.80 to 1.00. The opposite was observed for ploidy levels 18 and 20, as almost all of the loci (both in the biparental population and the panel) had medians in the range 0.40 to 0.60. Most of the individual medians at ploidy levels 12 and 14 were between 0.60 and 0.80, whereas the individual means at ploidy level 16 was evenly distributed in the ranges of 0.60 to 0.80 and 0.40 to 0.60.

Discussion

Developing a consistent and self-contained analysis depends on being able to estimate ploidy in species that have complicated genome structures, such as sugarcane. Due to its particular domestication process that involves the unequal participation of the parental species' genomes (*Saccharum officinarum* and *S. spontaneum*, known to have high chromosome numbers), cytogenetic studies may not be reliable under some circumstances. The approach used in this study combined mass spectrometry and the computer program SuperMASSA[13]

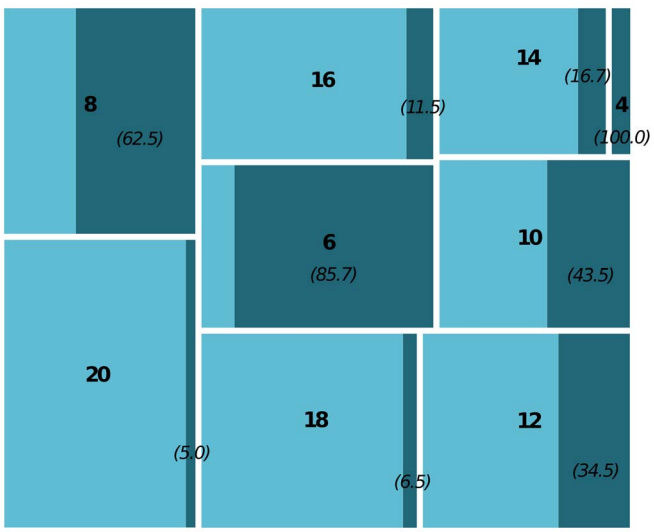

Figure 4 | Proportion of loci with a single dose in the biparental population. Loci were classified as single dosage when they had a SNP with only a single copy of one of the alleles in one parent, being a nulliplex in the other (thus segregating in a 1 : 1 fashion in the progeny), or when both parents had a single copy of the same allele (segregating in a 3 : 1 ratio). The areas of the rectangles are proportional to the number of SNPs of each ploidy level, indicated in bold font. SNPs with single doses are represented in dark blue, with the proportion of the respective ploidy in parenthesis. Only SNPs within categories A and B (see Figure 3) were considered.

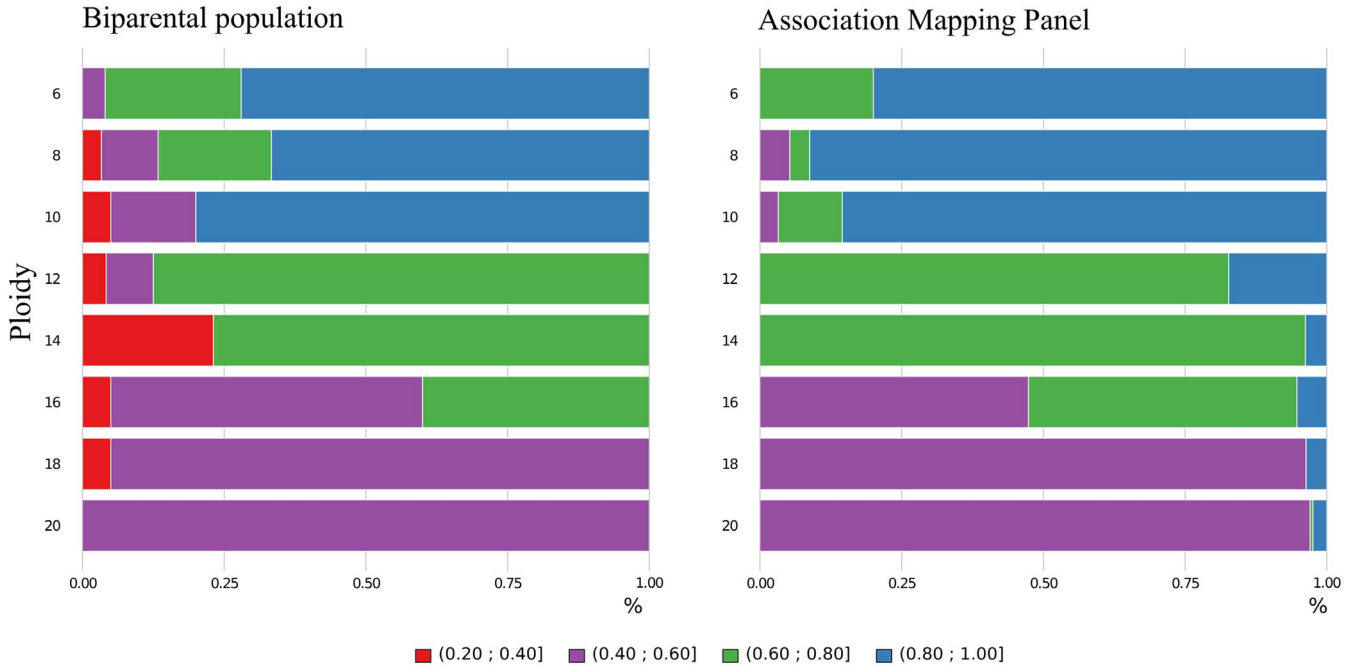

Figure 5 | Distribution of SuperMASSA individual posterior probabilities. For each locus, the median of all individual posterior probabilities was calculated. For instance, a median of 0.80 indicated that 50% of the individual posterior probabilities were greater than 0.80. The graphs show the distribution of the medians of each SNP locus that were classified with a specific ploidy. Only loci of category A (see Figure 3) were considered in this analysis.

and accomplished this task by simultaneously considering parental information, the number of clusters, the intensities associated with the different alleles, the expected frequencies of individuals in each cluster, and experimental error. During the analysis, each individual was assigned to a single cluster (genotype calling) with a high degree of confidence for several loci (Figures 3 and 5). Thus, we were able to use this technique to suggest a model to explain the complex genome structure of sugarcane.

The primary advantage of the approach used by SuperMASSA is that it makes use of the distribution of alleles in the population in addition to the relative intensities of each allele. Using both types of information is important for resolving cases in which similar relative allele intensities could be produced. For instance, tetraploid and octoploid individuals can both produce relative allele intensities of *0*:*4*, *1*:*3*, *2*:*4*, *3*:*1*, and *4*:*0*; however, if no distinct clusters of individuals with relative intensities near *1*:*7*, *3*:*5*, *5*:*3*, or *7*:*1* are observed, then it is highly unlikely that the population is octoploid. No two octoploid parents (or, if no parental data are considered, no Hardy-Weinberg allele frequency) can be expected to produce alternating observed and absent genotype classes. Because several ploidies can potentially produce clusters with similar relative allele dosages, exploiting population information is critical when inferring the ploidy level. In sugarcane, this advancement is particularly useful because the exact allele dosage of a locus is frequently unknown. Furthermore, the availability of parental data adds further constraints and increases the accuracy of ploidy estimation.

Most genetic studies of sugarcane have considered only simplex markers[3], and our results show that the actual portion of the genome explored to date is rather small. Our observations are quite different from previously published results. For example, one study reported that 80% of the AFLP markers in a biparental population occurred at a single dose[16]. This is similar to the values we found for ploidy levels from 6 to 12 for loci with category A, but not for the overall genome, suggesting that the strategy for finding single-dose loci may involve a biased genome sampling. Those authors considered only markers that segregated in only one parent, but there is no biological reason to support this approach because both parents can have different alleles segregating in the population[17,18]. However, it is important to note that AFLP analysis does not allow the identification of all genotypic classes in a segregating population because all clusters that have at least one copy of the allele will collapse into a single cluster (i.e., a dominant action). This also suggests that the identification of single-dose loci using AFLP is strongly biased.

What can be said about the sugarcane ploidy level? Our results suggest that the most likely ploidy levels are between 6 and 14 (Figure 5), and several lines of evidence support our findings. The genetic maps that have already been published using different sugarcane population types (e.g., biparental crosses, selfings, and others) all have recognised homo(eo)logy groups; interestingly, most homo(eo)logy groups were established with particular numbers of co-segregation groups, which also supports the mixed-ploidy nature of the sugarcane genome, consistent with the results presented here. Our estimates for ploidies 6–14 showed high (or intermediate) individual *posterior* probabilities. Furthermore, the proportion of loci with single dosages for these ploidy levels in the biparental population (Figure 4) is in agreement with previous reported results (e.g[16].), with the exception of ploidy 6. The proportion of loci with ploidy levels between 6 and 14 was approximately the same for loci within category A, both in the biparental population and in the genotype panel (Figure 3). This was expected because sugarcane chromosomes are approximately the same size and the markers were in principle chosen to evenly cover the genome. There is also biological evidence to support these findings; ploidies 6 to 14 are found in the group of species that contribute to the generation of modern cultivars of sugarcane. *S. officinarum* is the domesticated sugar-producing species that is directly derived from *S. robustum*, which encompasses clones with $2n = 60$ or $2n = 80$. Both species are autopolyploids, and their basic chromosome number is $x = 10$, meaning that *S. robustum* has 6 or 8 copies of each chromosome, depending on the genotype analysed[19–21]. A total of 13.4% of the SNPs used to genotype the biparental population and 7.12% of the SNPs used in the panel in this study have their level of ploidy classified as 6.

We speculate that this class of SNPs belongs to the subgenome (or haplotype) of *S. robustum* that persists in the sugarcane genotypes after breeding. The vast majority of *S. officinarum* clones display $2n = 80$ chromosomes. The species is stated to have eight sets (or copies) of 10 chromosomes ($x = 10$), i.e., octoploid.

Currently, it is supposed that modern sugarcane cultivars could exhibit $2n$ (*S. officinarum*) + n (*S. spontaneum*) constitution; when hybrids with *S. spontaneum* are produced, the chromosomes of *S. officinarum* double their number and form pairs of homologues, and those of *S. spontaneum* pair among themselves. This point was considered in classical publications[22,23]. Subsequent *in situ* hybridisation-based studies have confirmed the basic chromosome numbers (x) in the genus *Saccharum*[24] and suggested that the genomes of modern hybrids are composed of 10–20% *S. spontaneum* chromosomes, 5–17% recombinant chromosomes and the remainder composed of *S. officinarum* chromosomes[25,26]. Therefore, one would expect to find 8 as the most frequently estimated ploidy level, all derived from *S. officinarum*. This particular value was found in 26.7% of SNPs classified in Category A (considering only ploidies 6–14) in the biparental population (Figure 3a) and 10.1% SNPs used in the panel of genotypes and belonging to category A (Figure 3b). A possible explanation is that almost all genotypes analysed here were commercial varieties (mainly interspecific hybrids) with a modified chromosomal composition from the ancestors as a result of domestication.

For *S. spontaneum*, which displays a wide range of chromosome numbers (from $2n = 40$ to $2n = 128$), a basic chromosome number of $x = 8$ was suggested. The five major cytotypes with $2n = 64, 80, 96, 112$, and 128[27] have 8, 10, 12, 14 and 16 sets (or copies) of eight chromosomes, respectively. These are consistent with the values observed in this study. We may suppose that all these SNP-containing loci were inherited from *S. spontaneum* (maybe as haplotypes) or that they are located on the chromosomes that were identified as recombinants between the two species *S. officinarum* and *S. spontaneum*. Alternatively, when looking at ploidy level 8, all chromosomes could be inherited only from *S. officinarum*. It is also important to mention that the repeated cycles of backcrosses to *S. officinarum* applied by early breeders, combined with the double transmission phenomenon[22,23], could result in high ploidy levels because the contribution of the recurrent parent will be prevalent.

Chromosomal rearrangements are reported to be a rapid response to the formation of allopolyploid genomes[28]; intergenomic translocations occur predominantly between homo(eo)logous chromosomes[29], and homo(eo)logous shuffling and chromosome compensation maintain genome balance in re-synthesised allopolyploids[30]. All the rearrangements may have occurred in the early evolutionary process of modern sugarcanes. Supposedly, there is a most regular ploidy level, and all variations represent chromosome rearrangements that were herein observed.

The observation of 18 or 20 copies of a SNP-containing locus does not mean that this extreme figure represents the actual ploidy level. One could suggest reasonable cytological explanations for these high numbers; for example, for at least some of these loci, we may be detecting polysomic loci as a consequence of chromosomal segment copy number due to chromosomal rearrangements. On the other hand, the presence of univalents as a result of intergenomic pairing is well documented in sugarcane varieties. One should assume that bivalent pairing is not random but rather involves the same homo (eo)logous chromosomes[31]; therefore, two (or more) copies of the same univalent can be inherited from ascendants and pair during meiosis. The detection of certain high-copy SNP-containing loci may be a consequence of additional non-homologous pairing. However, it is important to mention that high values of ploidy were associated with some loci that did not have a reliable classification in our study. They were also more frequent in the panel, which is more difficult to analyse. Loci with ploidy 16 fall between these two scenarios (ploidy 6–14 and 18–20).

A recent review of the quantitative genotyping of polyploids[14] reported that, even when data are difficult to analyse (i.e., presenting high variance or strong allele-specific bias), the SuperMASSA software can still provide useful information to help to interpret the results and allow the evaluation of the reliability of those results. In that review, the authors evaluated the *posterior* probabilities of extremely high ploidies (in the range 2 to 100). It is obvious that most of these ploidies do not have biological support, but the study revealed that when the locus displays a high variance, the generated model tends to attribute a cluster to each point in the diffuse cloud, resulting in a very high estimate of ploidy level. We have not tried to adjust our models with ploidy levels above 20 due to computing-time limitations, but we have deliberately included ploidy values without biological support (2 and 4) or with weak evidence (18 and 20). The results show that this was a good strategy because the resulting individual *posterior* probabilities were rather small, indicating that our observations of high ploidy values (18 and 20) are likely to be explained as discussed above[18].

We have also performed some simulations to better understand the SuperMASSA output under normal and extreme situations (Figures S1 to S10, Supplementary Material). We observed that the software performed well when no extreme violations of its underlying assumptions were considered (for example, skew on the expected angles of clusters and segregation distortion). However, in the presence of high segregation distortion (for example, due to preferential pairing at meiosis) or some bias in the allele ratios, the estimated ploidy could be rather high (18 or 20).

The *in situ* hybridisations also helped us interpret the SuperMASSA estimations (Figure 6). The number of observed blocks (or signals) in these hybridisations could be taken as a rough estimate of ploidy level for IACSP 93-3046 (P2 of the biparental population). For *SugSNP382*, which yielded good and reliable results in the ploidy analysis (Figure 3), the number of observed blocks has been 8, which is close enough to the estimate of 10 provided by SuperMASSA. It is important to mention that SuperMASSA uses segregation ratios as an important feature to estimate ploidy; this is not necessarily the same as estimating ploidy by chromosome counting. For example, a homo(eo)logy group could have 10 chromosomes: 6 from *S. officinarum* and 4 from *S. spontaneum*. If there is preferential pairing at meiosis and the polymorphism is present in the genome of *S. officinarum*, the locus will behave like a hexaploid in the segregating population; in contrast, the results from cytological studies revealed a ploidy of 10. For *SugSNP715* and *SugSNP151*, the number of blocks was 10 for both. This is clear evidence that, as previously explained, high ploidy estimates (16 and 18) combined with small individual *posterior* probabilities are likely to be statistical artefacts. Moreover, *SugSNP382* yielded the same estimate for the ploidy level in the biparental population and the panel of genotypes, which was not the case for the other two SNPs analysed in Figure 3.

In conclusion, the results derived from the two different scenarios presented here (a biparental population and a panel of genotypes) provide extremely useful insights. First, as expected, it is clear that the sugarcane genome is complex and that the number of chromosomes in each hom(e)ology group varies depending on the SNP-containing locus. Second, our results agree with previous sugarcane cytogenetic data[25] and demonstrate the robustness of analysing SNP markers in autopolyploid species. Third, the ploidy level of each SNP locus was also estimated; it must be emphasised that this estimation cannot be performed with common marker systems.

In the light of our results, the ploidy of sugarcane commercial varieties (interspecific hybrids) was estimated to be in the range from 6 to 14 for each homo(eo)logy group; this has biological and statistical support. Several factors may explain the observation of estimates in the 16–20 ploidy range, a) they are actual results; b) they were caused by a combination of preferential pairing at meiosis and a lack of bivalent pairing or segregation distortion; c) there are intrinsic

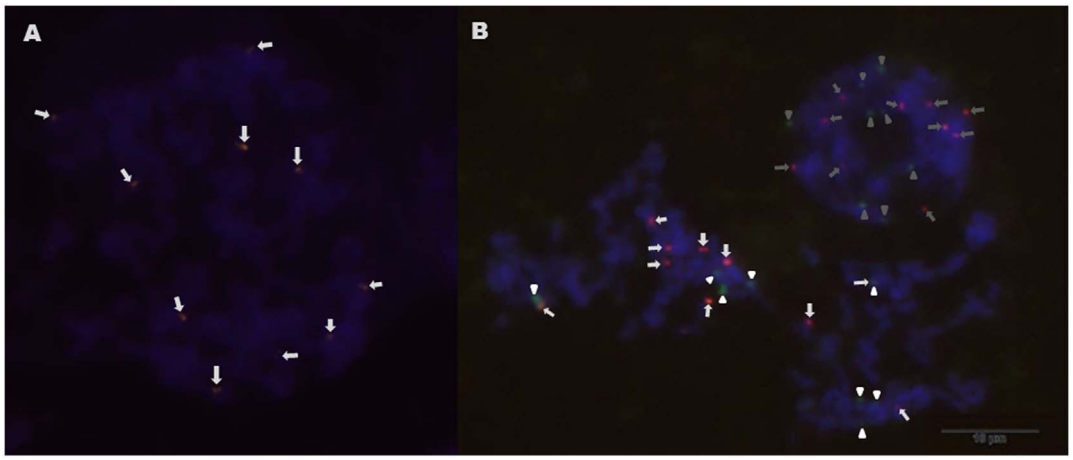

Figure 6 | *In situ* hybridisation of IACSP93-3046 chromosomes. (A) *SugSNP715* (10 blocks, arrows); (B) *SugSNP151* (10 blocks; grey arrow, nucleus; white arrow, metaphase nucleus) and *SugSNP382* (8 blocks; grey arrowhead, interphase nucleus; white arrowhead, metaphase nucleus).

difficulties in analysing loci with high ploidy and allelic dosage; or d) MassARRAY technology did not perform well for some loci, causing bias in the allele ratio and/or high variance for clusters. The results reported in the literature and our own *in situ* hybridisations for the three selected SNPs suggest that reason (a) is very unlikely. However, if these high estimates were actual results, further linkage studies will show that these loci with high ploidy will show evidence of linkage with other loci of the same ploidy level and, also, will not be linked with the ones in the ploidy range 6–14. It is important to mention that linkage studies based on genetic maps will require the development of new statistical approaches, such as the ones presented by[32] and[33] for autotetraploids, that would not be straightforward to use for our results. Current ideas that put strong emphasis on single-dose loci are not appropriate. Concerning point (b), this argument may be verified by further cytological information, which will help us understand the meiotic behaviour of this complex species and subsequently make modifications to the underlying assumptions in the statistical model. For explanations (c) and (d), specific procedures should be developed to optimise the methodology for dealing with complex polyploids. It is reasonable to assume that if most of these high ploidy values are true, these loci will co-segregate and will not be linked with loci with small ploidy; this will result in homo(eo)logous groups for the corresponding ploidy level. It is important to perform linkage studies in the biparental population to determine if loci with high/unknown ploidy are co-segregating with others showing high *posterior* probability for ploidy level; then the ploidy of these loci could be indirectly inferred.

None of the other currently available approaches are suitable to investigate polyploid genome structures as comprehensively as this approach. Therefore, we anticipate that the shaping of polyploid genomes by evolutionary processes will be better understood by applying this SNP genotyping method. Considering that most of the angiosperms are polyploid[34] and recent sequenced genomes also suggest a polyploid ancestry for eukaryotes[1], significant scientific breakthroughs can be achieved using this novel approach.

We strongly believe that the results presented herein will lead to new possibilities for the study of complicated autopolyploid species not only in terms of new genetic understanding, statistical genetic modelling, and prediction capabilities but also in terms of understanding the biological aspects of evolutionary and domestication processes. Finally, it is interesting to note that our study unveiled the genomic structure of a complex polyploid species by exploiting the simplest manifestation of genetic variation, the SNP. This approach should provide an important tool for developing high-quality genetic maps that will assist in QTL mapping and the assembly of reference genome sequences for the large proportion of plant plants species that are polyploid or have duplicated chromosomal regions.

Methods

Molecular and cytological analysis. Two representative scenarios were considered: a) a progeny of 180 individuals from a sugarcane F1 biparental population derived from the cross between two commercial varieties, IACSP 95-3018 (female, named P1 along the text) x IAC93-3046 (male, named P2); and b) an association mapping panel with 142 relevant sugarcane genotypes (Table 1), representing commercial varieties and important ancestors of modern cultivars. Sugarcane genomic DNA was obtained from young leaves using standard techniques. A total of 1034 sugarcane SNPs were developed; 91 were derived from previously reported sequence data[35] (Table S1), and the remaining 943 were developed from 2908 cluster sequences with differential expression[36] that were selected from the SUCEST database[37] (Table S2). SNPs were discovered using QualitySNP software[38] with minor modifications, and primers were designed using the MassARRAY Assay Design package. All 1034 sugarcane SNPs (Tables S1 and S2) were genotyped in the association mapping panel (iPLEX GOLD chemistry, Sequenom Inc., San Diego/CA, USA) (Table 1), and 271 SNPs from these (SUCEST database, Table S1) were evaluated in the progeny of the biparental population. Due to data quality control (especially due to very low signal), the data from 22 and 47 SNPs were discarded from the biparental population and from the panel of genotypes, respectively. Therefore, for the statistical analysis, 249 and 987 SNPs were used in the biparental population and in the panel, respectively. The SNP genotyping method was based on MALDI-TOF analysis performed on a mass spectrometer platform from Sequenom Inc.® Both parents from the biparental population were scored 12 times for each SNP.

The SNP assay is based on the single-base extension of locus-specific primers followed by mass spectrometry to detect polymorphisms, yielding allele-specific information[8]. Assuming equal ionisation efficiency for all alleles, equal PCR amplification of alternate alleles, and equal nucleotide incorporation accuracy/equilibria, the mass intensities should be proportional to the abundances of each allele.

Three selected SNPs (*SugSNP382*, *SugSNP151*, and *SugSNP715*) were analysed with FISH to check their hybridisation with IACSP93-3046. Leaf genomic DNA was isolated using the DNeasy Plant Mini Kit (Qiagen) and amplified using a *Pfu* DNA Polymerase kit (Thermo Scientific) and specific primers (Table S3). The fragments of DNA were cloned using *Escherichia coli* DH10b as host and pGEM-T Vector Systems (Promega) as vector. Colonies containing recombinant plasmids were identified for selection on LB agar medium supplemented with X-gal and IPTG. Recombinant plasmids were isolated using the alkaline minipreparation procedure, and the insert nucleotide sequences were determined with an ABI3500 automated DNA sequencer (Applied Biosystems). DNA Sequences were analysed with Lasergene 7 (DNAStar, Madison, WI, USA) and aligned by using the ClustalW option of the MegAlign program. The clones were used to amplify the probes for FISH using *Taq* DNA polymerase (Invitrogen) and purified using Wizard® SV Gel and PCR Clean-Up System (Promega). Chromosome preparations were made from root tips collected from culms grown in a plastic box containing filter paper with the regular application of water. Cytological preparations were carried out as previously described[39]. All probes were labelled by nick translation (Invitrogen). *SugSNP715* and *SugSNP151* were labelled with digoxigenin-11-dUTP (Life Technologies) and detected with Anti-DIG-rhodamine; *SugSNP382* was labelled with Biotin-14dATP (Roche) and detect with avidin-FITC. The procedure and conditions for FISH were previously described[40].

Statistical analysis. The output of Sequenom iPLEX MassARRAY technology is a scatter plot D with quantitative alleles intensities for individuals $i = 1, 2$, up to n[8,13]. Because each SNP was bi-allelic, two intensities are presented, x_i and y_i, usually

Table 1 | Genotypes from the panel of 142 sugarcane varieties (panel of genotypes)

Badila	IAC87-3396	RB735275	RB92579
CB36-24	IAC91-1099	RB739359	RB935744
CB40-13	IACSP93-2060	RB739735	RB965902
CB41-76	IACSP93-3046	RB75126	RB965917
CB46-47	IACSP95-3018	RB765418	RB966928
CB47-355	IACSP95-3028	RB785148	SP70-1005
CB53-98	IACSP95-5000	RB815690	SP70-1078
Chunnee	IACSP98-3022	RB825317	SP70-1143
Co290	IN84-58	RB825336	SP70-1284
Co331	L60-14	RB83102	SP70-1423
Co419	Maneria	RB835019	SP70-3370
Co449	NA56-79	RB835054	SP71-1406
Co740	NCo310	RB835089	SP71-6163
Co997	PO88-62	RB835205	SP71-6949
CP51-22	Q165	RB835486	SP71-799
CP52-68	R570	RB845197	SP72-4928
CP70-1547	RB1	RB845210	SP77-5181
CTC15	RB2	RB845257	SP79-1011
CTC2	RB3	RB855002	SP79-2233
CTC9	RB4	RB855035	SP79-2312
EK28	RB5	RB855036	SP79-2313
F31-962	RB6	RB855077	SP79-6134
F36-819	RB7	RB855113	SP79-6192
Ganda Cheni	RB8	RB855156	SP80-1520
H53-3989	RB9	RB855206	SP80-1816
H59-1966	RB10	RB855350	SP80-1836
IAC48-65	RB11	RB855453	SP80-1842
IAC49-131	RB12	RB855463	SP80-3280
IAC50-134	RB721012	RB855511	SP81-3250
IAC51-205	RB72199	RB855536	SP83-2847
IAC52-150	RB72454	RB855563	SP83-5073
IAC64-257	RB725053	RB855595	SP89-1115
IAC82-2045	RB725828	RB867515	SP91-1049
IAC82-3092	RB732577	RB925211	TUC71-7
IAC83-4157	RB735200	RB925268	
IAC86-2210	RB735220	RB925345	

represented in bi-dimensional scatter plots (see Figure 1 for an example of a loci with alleles T and G). For data quality, all data points with small intensities for both alleles were removed; they were located within a circular area on the scatter plots defined by the radius $(0.10)min\{x_i, y_i\}$, centred on the origin of both axes.

All loci were then classified using the statistical method implemented in the SuperMASSA software[13]. A comprehensive review of this method is presented in[14]. In short, rather than iteratively clustering the samples and then predicting the genotype of each cluster, a graphical Bayesian method was used. The model can be described in two parts. First, a Gaussian model based on the relative dosage is used to model the probability that an individual with a known genotype will produce certain intensities for each allele; ideally, the relative intensities would be proportional to the relative dosages of the respective alleles. Second, a multinomial distribution is used to model the probability that a given set of genotypes will occur given the population structure. The population structure is general and can be used to analyse the biparental population (F1 model) and the association mapping panel (Hardy-Weinberg model). For any type of population model, the hidden parameters (i.e., the allele frequency for the Hardy-Weinberg model and the parental genotypes in the F1 model) can be estimated with maximum likelihood. Similarly, the ploidy can be predicted by estimating the genotypes and population parameters for each ploidy level and then selecting the ploidy that yields the highest likelihood. In the case of the F1 model, additional data were provided by the parents, which were scored with 12 replicates; these data can help restrict the set of reasonable parents and ploidies. The primary contribution of this method is that it makes use of the distribution of alleles in the population and the relative intensity of each allele. The use of both types of data is important for resolving cases that could produce similar relative allele intensities.

Following the recommendation reported in[14], to find the *maximum a posteriori* (MAP) solution for the estimates of the parameters in the model, all even-numbered ploidy levels in the range of 2 to 20 were tested. The SuperMASSA *naive posterior report threshold* was set to 0, and the values of individual *posterior* probability (which indicates the maximum threshold that will allow the individual to be assigned to a given genotype) were also calculated. For example, if two individuals have *posterior* probabilities 0.55 and 0.65 and the *naive posterior report threshold* is set to 0, both of them will be assigned to genotypes; changing the threshold to 0.60, only the latter will be included; with a threshold of 0.90, both will be excluded. This was shown to be important when interpreting the results of the SNP calling.

1. Comai, L. The advantages and disadvantages of being polyploid. *Nature Rev. Genet.* **6**, 836–846 (2005).
2. Alwala, S. & Kimbeng, C. A. *Genetics, Genomics and Breeding of Sugarcane.* Henry, R. J. & Kole, C. (ed.), 69–96 (CRC Press, 2010).
3. Pastina, M. M., Pinto, L. R., Oliveira, K. M., Souza, A. P. & Garcia, A. A. F. *Genetics, Genomics and Breeding of Sugarcane*. Henry, R. J. & Kole, C. (ed.) 117–148 (CRC Press, 2010).
4. Cao, D., Craig, B. A. & Doerge, R. W. A model selection-based interval-mapping method for autopolyploids. *Genetics* **169**, 2371–2382 (2005).
5. Doerge, R. W. & Craig, B. A. Model selection for quantitative trait locus analysis in polyploids. *Proc. Natl. Acad. Sci USA* **97**, 7951–7956 (2000).
6. Galitski, T., Saldanha, A. J., Styles, C. A., Lander, E. S. & Fink, G. R. Ploidy regulation of gene expression. *Science* **285**, 251–254 (1999).
7. Osborn, T. C. *et al*. Understanding mechanisms of novel gene expression in polyploids. *TIG* **19**, 141–147 (2003).
8. Gabriel, S., Ziaugra, L. & Tabbaa, D. SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Curr. Protoc. Hum. Genet.* **60**, 2–12 (2009).
9. Akhunov, E., Nicolet, C. & Dvorak, J. Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. *Theor. Appl. Genet.* **119**, 507–17 (2009).
10. Elshire, R. J. *et al*. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6**, e19379 (2011).
11. Baird, N. *et al*. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS ONE* **3**, e3376 (2008).
12. Liu. *et al*. High-Throughput Genetic Mapping of Mutants via Quantitative Single Nucleotide Polymorphism Typing. *Genetics* **184**, 19–26 (2010).
13. Serang, O., Mollinari, M. & Garcia, A. A. F. Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. *PloS ONE* **7**, e30906 (2012).
14. Mollinari, M. & Serang, O. Quantitative SNP Genotyping of Polyploids with MassARRAY and Other Platforms. *Methods in Molecular Biology*. Walker, J. M. (ed.) (Humana Press, 2013 -in press).
15. Grivet, L. & Arruda, P. Sugarcane genomics: depicting the complex genome of an important tropical crop. *Curr. Opin. Plant Biol.* **5**, 122–127 (2001).
16. George, A. W. & Aitken, K. A new approach for copy number estimation in polyploids. *J. Hered.* **101**, 521–524 (2010).
17. Garcia, A. A. F. *et al*. Development of an integrated genetic map of a sugarcane (*Saccharum spp*.) commercial cross, based on a maximum-likelihood approach for estimation of linkage and linkage phases. *Theor. Appl. Genet.* **112**, 298–314 (2006).
18. Oliveira, K. M. *et al*. Functional integrated genetic linkage map based on EST-markers for a sugarcane (*Saccharum spp*.) commercial cross. *Mol. Breeding* **20**, 189–208 (2007).
19. Price, S. Cytology of *Saccharum robustum* and related sympatric species and natural hybirds. *USDA Tech. Bull.* **1337**, 1–44 (1965).
20. Daniels, J. & Roach, B. T. *Sugarcane Improvement Through Breeding*. Heinz, D.J. (ed.) (Elsevier, Amsterdam, 1987).
21. Lu, Y. H., D'Hont, A., Walker, D. I. T. & Rao, P. S. Relationships among ancestral species of sugarcane revealed with RFLP using single copy maize nuclear probes. *Euphytica* **78**, 7–18 (1994).
22. Brandes, E. W. & Sartoris, G. B. Sugarcane: Its Origin and Improvement. *Yearbook of the United States Department of Agriculture*, 561–623 (USDA, 1936).
23. Bremer, G. Problems in breeding and cytology of sugar cane. *Euphytica* **10**, 59–78 (1961).
24. D'Hont, A., Ison, D., Alix, K., Roux, C. & Glaszmann, J. C. Determination of basic chromosome numbers in the genus *Saccharum* by physical mapping of ribosomal RNA genes. *Genome* **41**, 221–225 (1998).
25. D'Hont, A. *et al*. Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum spp*.) by molecular cytogenetics. *Mol. Gen. Genet.* **250**, 405–413 (1996).
26. D'Hont, A. Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. *Cytogenet. Genome Res.* **109**, 27–33 (2005).
27. Panje, R. & Babu, C. Studies in *Saccharum sponteneum*. Distribution and geographical association of chromosome numbers. *Cytologia* **25**, 152–172 (1960).
28. Pontes, O. *et al*. Chromosomal locus rearrangements are a rapid response to formation of the allotetraploid Arabidopsis suecica genome. *P. Natl. Acad. Sci. USA* **101**, 18240–18245 (2004).
29. Chester, M. *et al*. Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). *P. Natl. Acad. Sci. USA* **109**, 1176–1181 (2012).
30. Xiong, Z., Gaeta, R. T. & Pires, J. C. Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*. *P. Natl. Acad. Sci. USA* **108**, 7908–7913 (2011).
31. Pagliarini, M. S., Vieira, M. L. C. & Valle, C. *Meiosis – Molecular Mechanisms and Cytogenetic Diversity.* Swan, A. (ed.), (InTech, 2012).
32. Leach, L. J., Wang, L., Kearsey, M. J. & Luo, Z. Multilocus tetrasomic linkage analysis using hidden Markov chain model. *P. Natl. Acad. Sci. USA* **107**, 4270–4274 (2010).
33. Hackett, C. A., McLean, K. & Bryan, G. J. Linkage Analysis and QTL Mapping Using SNP Dosage Data in a Tetraploid Potato Mapping Population. *PLoS ONE* **8**, e63939 (2013).

34. Masterson, J. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science* **264**, 421–424 (1994).
35. Bundock, P. C. *et al*. Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploidy plant species using 454 sequencing. *Plant Biotech. J.* **7**, 347–354 (2009).
36. Papini-Terzi, F. S. *et al*. Sugarcane genes associated with sucrose content. *BMC Genomics* **10**, 120 (2009).
37. Vettore, A. L. *et al*. Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Res.* **13**, 2725–2735 (2003).
38. Tang, J., Vosman, B., Voorrips, R. E., Gerard van der Linden, C. & Leunissen, J. A. M. QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploidy species. *BMC Bioinformatics* **7**, 438 (2006).
39. Moraes, A. P. & Guerra, M. Cytological differentiation between the two subgenomes of the tetraploid *Emilia fosbergii* Nicolson and its relationship with *E. sonchifolia* (L.) DC. Asteraceae). *Plant Syst. Evol.* **287**, 113–118 (2010).
40. Schwarzacher, T. & Helslop-Harrison, J. S. *Pratical in situ Hybridization*. (BIOS Scientific Publishers, 2000).

Acknowledgments

The authors wish to thank Dr Daniel Ugarte for his invaluable suggestions for elaboration of the manuscript. This work was supported by grants from Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) - Bioen Program, grant no. 2008/52197-4, and Conselho Nacional do Desenvolvimento Científico e Tecnológico (CNPq) – INCT- Bioetanol and CeProBio Project. CBCS, DAS, EAC, MCM, MM, and MMP received graduate fellowships from FAPESP. TGM and RRS received a fellowship from CNPq. ND, MOSG, and RG received post-doctoral fellowships from FAPESP. AAFG, APS, ERFM, GMS, MLCV, MAVL, MV, and RVe received research fellowships from CNPq.

Author contributions

R.Vi. and T.G.M. identified and developed the sugarcane SNPs. A.A.F.G., M.M., M.M.P., O.R.S., R.G. and R.R.S. performed the statistical analysis. M.O.S.G. and T.G.M. carried out the SNP genotyping. L.R.P. and M.G.A.L. carried out the biparental cross. L.R.P., M.G.A.L. and M.S. were responsible for the field trials and the collection of the plant material. E.A.C., C.B.C.S., G.M.S., L.R.P., M.A.V.S., M.C.M., M.S.C., M.V., P.B., R.H. and R.Ve. participated in the molecular genetic studies. E.R.F.M., N.D. and D.A.S. performed *in situ* hybridisation experiments. M.L.C.V. provided the cytogenetic interpretation of the results. A.A.F.G. and A.P.S. conceived the study and participated in its design and coordination. A.A.F.G., M.M. and A.P.S. drafted the manuscript. All authors read and approved the final manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Garcia, A.A.F. *et al*. SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. *Sci. Rep.* **3**, 3399; DOI:10.1038/srep03399 (2013).